Group 5 Write Up

In this lab, we were asked to create a spam filter using a data set from the 2007 TREC Public Spam Corpus. Using this information, our group explored and experimented with different model hyperparameters for our spam filter. We found that by tweaking a few of the parameters gave very different results. We first started out by executing the code in file spamFilterLogisticRegression.py and changing the training to a better ratio of 80:20, this led to greater precision and accuracy than the original. We then changed the code up multiple times again with new classifications. We ran the code again using a decision tree, support vector machine, random forest classifier and an ANN classifier. After running all of these individually we then compared performance metrics. Looking at our performance metrics from our tests we can see that support vector machine was the best for obtaining the greatest results with an overall Precision of 0.9924205773262377. We have complied all our python code and performance metrics for each test into a zip for further analysis.